中图法分类号:TP18; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-17

论文引用格式:Cai Weinan, Wang Zongji, Zhang Yuanben, Yin Yuhao, Liu Junyi. Generative Extrapolation of Sparse Aerial Views Guided by 3D Priors[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260160. (蔡伟南, 王宗继, 张源奔, 殷煜昊, 刘俊义. 三维先验引导的稀疏航拍视角生成式外推[J/OL]. 中国图象图形学报,XXXX:1-17. DOI: 10.11834/jig.260160.)[DOI:10.11834/jig.260160]

三维先验引导的稀疏航拍视角生成式外推

蔡伟南[1,2,3],王宗继[1,2],张源奔[1,2],殷煜昊[1,2,3],刘俊义[1,2]

1. 中国科学院空天信息创新研究院,北京　100094; 2. 目标识别与应用技术重点实验室,北京　100094; 3. 中国科学院大学,北京　101408

摘　要:**目的**　在当前工业级的无人机航测与实景三维重建工作流中,利用摄影测量或机载 LiDAR 快速获取测区的稀疏三维点云已成为标准工序。然而,受限于无人机续航、空域管制或既定航线规划,实际采集的光学影像往往难以实现全视角的密集覆盖。当利用这些离散且带有盲区的三维先验进行大基线视角外推或传统图形学渲染时,画面极易产生严重的结构畸变与渲染伪影。另一方面,纯数据驱动的二维扩散模型由于缺乏三维物理约束,在大基线视角外推时,纯二维模型极易打破极线几何约束,导致严重的透视畸变与地物拓扑错位。**方法**　为突破这些问题,本文提出一种融合稀疏三维先验的生成式视角外推方法,旨在利用离散的物理几何骨架引导连续的像素生成,实现无人机稀疏视角下的高保真、几何一致性受控外推。该框架将多源物理空间约束显式注入潜在扩散模型(Stable Diffusion XL,SDXL),核心包含三个阶段:首先是三维先验的空间对齐。将点云提供的绝对深度与相机外参融合,构建像素级"深度+坐标"的几何空间嵌入图,并通过位姿变换将源影像预先对齐。其次是语义与几何解耦的双路前向生成。利用 IP-Adapter 提取源图像全局语义,同时通过 ControlNet 注入几何空间嵌入图联合特征,指导网络在保持精确空间拓扑的基础上,渲染出连续且逼真的高频纹理。最后是基于潜空间重投影的三维透视监督。利用单步去噪估计推算无噪特征,并在潜空间计算跨视角几何重投影损失,强制生成的纹理严格服从极线几何的物理透视规律。**结果**　基于真实无人机航拍及其解算点云数据集的实验表明,本方法在稀疏视角的大跨度外推任务中显著优于现有基线模型,有效抑制了生成过程中的透视畸变与地物拓扑错位。与传统的重建相比,本方法生成的图像在视觉感知上更为自然真实。测试结果显示,感知评价指标 LPIPS 降至 0.466,在 LLaVA-IQA 与 CLIP-Score 等语义一致性评估中也体现出了相应的优势。**结论**　综上所述,本研究为稀疏条件下的新视角合成探索了一种新的可行思路。我们打破了对传统多视图连续性的依赖,充分发挥了扩散模型的生成潜力。

关键词: 新视角合成;多视角一致性;几何空间嵌入;无人机航拍影像;扩散模型

Generative Extrapolation of Sparse Aerial Views Guided by 3D Priors

Cai Weinan[1,2,3], Wang Zongji[1,2], Zhang Yuanben[1,2], Yin Yuhao[1,2,3], Liu Junyi[1,2]

1. Aerospace Information Research Institute of Chinese Academy of Sciences; , Beijing 100094, China; 2. Key Laboratory of Target Cognition and Application Technology; , Beijing 101408, China; 3. University of Chinese Academy of Sciences, Beijing 101408

Abstract: **Objective** In the contemporary landscape of geospatial engineering and computer vision, modern industrial workflows dedicated to Unmanned Aerial Vehicle (UAV) aerial surveying and high-fidelity 3D scene reconstruction have become foundational. Rapidly acquiring sparse 3D point clouds of target areas—typically achieved through Structure-from-Motion (SfM) photogrammetry or airborne Light Detection and Ranging (LiDAR)—has been established as a standard procedure. However, practical data collection is frequently hindered by severe physical and regulatory bottlenecks. Restricted by limited UAV battery endurance, strict airspace control regulations, or rigid predefined flight routes, the actual optical

收稿日期:2026-03-28;**修回日期**:2026-04-25

imagery collected during field operations can rarely achieve the dense, omnidirectional overlapping coverage required by ideal reconstruction algorithms. Consequently, researchers are often forced to work with highly sparse visual observations. When attempting to perform large-baseline view extrapolation or traditional graphics rendering utilizing these discrete and occluded 3D priors, existing methods face insurmountable challenges. Traditional rendering pipelines often result in severe structural distortions, projection holes, and rendering artifacts due to the lack of dense multi-view continuity. On the other hand, the recent emergence of purely data-driven 2D generative diffusion models (such as Stable Diffusion) offers unprecedented capabilities in high-frequency texture hallucination. Yet, these models fundamentally lack explicit 3D physical constraints. During large-baseline view extrapolation, they easily violate epipolar geometric constraints, leading to severe perspective distortion, physical scale collapse, and the topological misalignment of critical ground objects. To address these critical limitations, this research proposes a novel generative view extrapolation method fused with sparse 3D priors. The primary objective is to utilize discrete physical geometric skeletons to constrain and guide continuous pixel generation. By bridging the gap between graphical rendering and generative modeling, this research aims to enable high-fidelity, geometrically consistent, and controllable novel view extrapolation under extreme, sparse UAV observation conditions. **Method** To achieve the aforementioned objectives, the proposed framework explicitly injects multi-source physical spatial constraints into a large-scale Latent Diffusion Model (SDXL), fundamentally transforming it from a pure 2D image generator into a 3D-aware extrapolation engine. The methodology is meticulously structured into three core stages to ensure that geometric consistency and semantic richness are simultaneously preserved. First, the framework executes the **Spatial Alignment of 3D Priors**. Recognizing that raw point clouds are discrete and challenging for 2D networks to interpret, we project the sparse point cloud onto the target camera view using the corresponding extrinsic matrices. This projection yields an absolute depth map. By fusing this depth information with pixel-level coordinate maps, we construct a comprehensive Geometric Spatial Embedding (GSE) map. Furthermore, to provide the generative network with an optimal starting initialization, the source reference image undergoes a pre-alignment process via rigid pose transformation, mapping the source pixels to the approximate target perspective based on available geometric data. Second, we propose a **Dual-Branch Forward Generation Pipeline with Decoupled Semantics and Geometry**. To prevent the network from confusing structural boundaries with texture colors, the architecture handles these modalities independently. An IP-Adapter (Image Prompt Adapter) is employed as the semantic branch to extract global appearance characteristics—such as lighting conditions, material textures, and environmental atmosphere—directly from the source image. Simultaneously, a ControlNet architecture serves as the geometric branch, meticulously injecting the constructed GSE features into the denoising steps. This decoupled dual-branch design forces the network to render continuous and realistic high-frequency textures (guided by the semantic branch) while strictly anchoring them to precise spatial topologies (constrained by the geometric branch). Third, the framework introduces **3D Perspective Supervision via Latent-Space Reprojection**. Traditional pixel-space losses often fail to capture deep structural semantic errors during the diffusion process. Therefore, our method computes the geometric constraints directly within the latent space. At each denoising timestep, the model estimates the noise-free latent features. Using the known camera intrinsic and extrinsic matrices, these latent features are reprojected across views to compute a cross-view geometric reprojection loss. This rigorous supervision mechanism continuously penalizes any spatial drift during the reverse diffusion process, forcing the newly generated textures to strictly follow the physical perspective rules of epipolar geometry. **Result** The proposed framework was rigorously evaluated using datasets derived from real-world UAV aerial imagery and their corresponding reconstructed point clouds. Our experimental setup was explicitly designed to simulate extreme large-baseline view extrapolation tasks, where the target camera pose significantly deviates from the available sparse reference views. We conducted comprehensive qualitative and quantitative comparisons against state-of-the-art baselines, including purely data-driven models like VistaDream and adapted baselines such as SDXL equipped with standard ControlNet depth conditioning. Qualitative visual analyses demonstrate that the proposed method significantly outperforms all existing baseline models. While foundational SDXL models suffered from severe structural drift—often manifesting as fractured roads or the unnatural intertwining of distinct architectural elements—our method accurately preserved the targeted ground objects. The generated objects exhibited geometrically correct recession strictly along the depth axis, maintaining precise topological boundaries without perspective distortion. Furthermore, compared to traditional multi-view ste-

reo reconstruction pipelines, the images synthesized by our generative framework eliminated rendering empty holes and appeared significantly more natural and visually realistic. Quantitatively, the proposed method achieved superior performance across multiple core perception and semantic metrics. Traditional pixel-level metrics like PSNR often penalize generative models due to minor, physically plausible texture variations. Therefore, we focused on metrics that reflect human visual perception and deep semantic alignment. Our method successfully reduced the Learned Perceptual Image Patch Similarity (LPIPS) score to an exceptional 0.466, objectively proving its capability to synthesize high-fidelity textures that closely mimic real UAV footage. Moreover, the framework exhibited distinct advantages in large-scale visual-language model evaluations, comprehensively outperforming baseline models in both LLaVA-IQA aesthetic scoring and CLIP-Score semantic consistency evaluation. **Conclusion** This research represents a significant paradigm shift in the field of novel view synthesis, moving away from relying on dense pixel interpolation towards geometry-guided generative extrapolation. By seamlessly integrating the deterministic "rendering" principles of traditional computer graphics with the "generative" capabilities of advanced diffusion models, we have explored a highly effective solution for large-baseline view synthesis under sparse observation conditions. The proposed method successfully breaks the long-standing dependence on multi-view continuity. By treating discrete 3D point clouds not merely as rendering primitives, but as rigorous physical anchors to constrain the generative diffusion process, the framework effectively mitigates spatial drift and perspective distortion. From a practical engineering standpoint, this research directly addresses the severe objective constraints faced in actual UAV aerial surveying, such as airspace restrictions and limited flight endurance. It provides a robust, controllable, and highly practical algorithm pipeline capable of accurately deducing large-scale spatial structures, thereby unlocking new potentials for digital twin modeling and complex scene reconstruction in industrial applications.

Key words: Novel View Synthesis; Multi-view Geometric Consistency; Geometric Spatial Embedding; Unmanned Aerial Vehicle Imagery; Latent Diffusion Models

0 引　言

0.1 研究背景

随着实景三维、数字孪生等技术的加速发展，无人机摄影测量与三维重建技术应用广泛。目前，机载 LiDAR 或低精度 DEM 已能高效获取测区宏观三维骨架，但受无人机续航、空域管制及气象条件限制，高分辨率光学纹理影像的获取成本仍然高昂（黄洋，郭宇等，2025）。在稀疏光学影像覆盖下，传统点云 CG 渲染技术在大基线视角切换时易出现结构拉伸与黑色空洞，难以满足高保真三维重建与漫游需求。因此，基于基础物理高程和极少量参考影像补全缺失视角的真实光学纹理，已成为遥感测绘与计算机视觉交叉领域的重要研究方向。

与此同时，以潜在扩散模型（Latent Diffusion Model，LDMs）为代表的生成式 AI 在图像合成与特征补全领域展现出强大的先验拟合能力（Rombach，Blattmann 等，2022），为稀疏视角纹理生成提供了新思路。但现有视觉生成大模型本质是纯数据驱动的二维概率分布学习，缺乏三维到二维的严格物理光学透视映射机制（Podell，English 等，2023）。直接应用于无人机六自由度（6-DOF）大尺度航拍场景时，因缺少显式极线几何先验约束，网络难以构建可靠的多视角空间拓扑关系，导致连续生成中出现严重的尺度歧义与像素“漂移”（Bernal-Berdun，Serrano 等，2025）。因此，打破纯数据驱动模型的二维生成平滑化倾向，引入绝对物理空间锚定以抑制空间漂移、实现高保真遮挡纹理补全，是该技术落地测绘应用的核心瓶颈。针对这一挑战，学术界在新视角合成与三维内容生成领域已开展广泛探索，形成多条技术演进路线。

0.2 相关工作

针对上述挑战，学术界在新视角合成（Novel View Synthesis，NVS）与三维内容生成领域开展了广泛探索，当前主流技术路线均以扩散模型为核心基础。扩散模型是一类基于概率的生成模型，其核心思想是通过学习逆向去噪过程，从高斯噪声中逐步恢复出真实数据分布（占瑞乙，樊轶等，2026）。为解决高分辨率像素空间计算成本过高的问题，潜在扩散模型（Latent Diffusion Models，LDMs）引入预训练的感知压缩模型，将图像映射到低维潜空间进行加噪与去噪操作，在保留核心语义信息的同时大幅降低了计算复杂度。其中 Stable Diffusion XL

(SDXL)作为目前最先进的LDM架构之一,通过扩大UNet主干规模、采用双文本编码器策略及多分辨率训练机制,显著提升了高保真、高分辨率图像的生成能力,成为后续各类三维生成方法的主流基础模型。根据生成逻辑与三维表征结合方式的差异,基于扩散模型的主流方法可归纳为四大类:基于得分蒸馏采样的优化方法(Optimization-based via SDS)、显式多视角协同扩散模型(Multi-view Diffusion Models)、前馈式大规模重建模型(Large Reconstruction Models, LRMs),以及基于视频生成先验的动态合成方法(Video-based NVS)。

0.2.1 基于得分蒸馏采样的优化方法

基于得分蒸馏采样的方法是二维图像生成向三维内容创制迁移的重要范式。其核心是避开大规模三维标注数据匮乏的困境,将预训练二维扩散模型的海量视觉先验"蒸馏"至三维表征中;该类方法不直接训练三维生成器,而是迭代优化NeRF(Mildenhall, Srinivasan等, 2020)、DMTet(Shen, Gao等, 2021)或3DGS(Kerbl, Kopanas等,2023)进行迭代优化。DreamFusion(Poole, Jain等,2022)首创SDS损失函数,实现文本驱动的NeRF生成;Magic3D(Lin, Gao等,2023)提出"粗到精"两阶段策略,解决了前者分辨率低、生成慢的问题;ProlificDreamer(Wang, Lu等,2023)进一步提出变分得分蒸馏,加速生成并提升了几何保真度。

尽管SDS方法在创意生成领域表现优异,但工业级应用仍存在多重瓶颈:一是效率低下,单物体生成需数十分钟,无法满足实时性要求;二是存在"多脸"(Janus)问题(Li, Zhang等,2024),二维先验缺乏三维几何常识,易导致复杂结构视角逻辑断层;三是SDS损失倾向于分布众数,生成纹理过度平滑且色彩过饱和。在无人机影像生成任务中,SDS方法随机性过高,难以精准复刻地表真实拓扑结构,其"幻觉"特性还易导致道路、建筑等地理信息严重失真,无法满足测绘级高保真合成要求。

0.2.2 显式多视角协同扩散模型

为解决SDS优化中的几何不一致性,研究重点转向原生多视角优化的扩散模型。此类方法通过修改视觉Transformer的注意力机制,使其生成时感知多视角信息,从而输出具有一致性的结果。Zero-1-to-3(Liu, Wu等,2023)首次将相对相机姿态显式注入Stable Diffusion,实现单图到新视角的转换;MVDream(Shi, Wang等,2023)提出三维自注意力机制,强制跨视角特征交换,显著缓解了多面问题;SyncDreamer(Liu, Lin等,2024)引入同步多视角扩散与体素导向,保持了采样过程的特征对齐;Wonder3D(Long, Guo等,2024)和Consistent-1-to-3(Ye, Wang等,2024)则进一步联合生成法线图与深度图,通过同步输出几何先验增强空间连贯性。

然而,这类方法在无人机数据上存在显著局限:一是现有模型多基于Objaverse(Deitke, Schwenk等,2023)等物体级数据集训练,具有强烈的"以物体为中心"归纳偏置,难以泛化至无人机"以场景为中心"的大尺度视角;二是模型通常假设简单相机分布,无法适配无人机影像的剧烈尺度变化与复杂6-DOF轨迹,大倾角或远距离视角下生成效果急剧下降;三是显式多视角模型对输入分辨率限制较严,难以直接处理5K+的超高分辨率倾斜摄影图像。

0.2.3 前馈式大规模重建模型

受大语言模型(LLM)启发,前馈式大规模重建模型(LRM)(Hong, Zhang等,2023)展现了"大数据+大模型"直接预测三维结构的潜力。该类方法摒弃逐场景优化,通过百万级3D数据预训练Transformer,学习稀疏视角到三维几何的强映射,实现快速推理。LRM采用三平面(Tri-plane)表示(Chan, Lin等,2022)在5秒内完成单图建模。LGM(Large Gaussian Model)(Tang, Chen等,2025)将输出转为三维高斯泼溅(3DGS),依托其高效渲染特性实现极速训练推理,提升了高光与细节表现;InstantMesh(Xu, Cheng等,2024)结合多视图扩散与稀疏重建器优势,可从单图生成网格结构。

但是LRM类方法虽速度极快,但在无人机场景中面临"精度与显存"的核心权衡:一是三平面或Token数量限制导致纹理分辨率不足,难以达到遥感级清晰度,易丢失地物边缘高频细节;二是LGM等方法对输入视角一致性高度敏感,参考图微小误差即可引发几何伪影或空洞;三是模型参数量大、训练微调成本高,且主要基于合成物体数据训练,直接迁移至真实大规模地形场景时会出现严重域偏移。

0.2.4 基于视频生成先验的动态合成方法

基于视频生成先验的方法利用视频扩散模型的时间连续性约束三维空间一致性,核心是将SVD(Jiang, He等,2019)等大规模视频数据集中学到的物理与几何先验迁移至3D生成。SV3D(Voleti,

Yao等，2025）基于Stable Video Diffusion（Blattmann，Dockhorn等，2023）微调，可输出特定轨道上的高一致性视频序列，解决了单图生成的几何塌陷问题；V3D（Chen，Wang等，2025）将视频模型作为密集视角采样器，通过高帧率生成显著提升了纹理细腻度；ConsistNet（Yang，Cheng等，2024）等进一步通过稀疏几何约束稳定了视频生成的空间位置。

但将视频先验应用于无人机数据合成仍面临多重挑战：一是轨迹控制难题，现有模型多假设平滑环绕轨道，难以适配无人机直飞、拉升、侧转等复杂非结构化轨迹，无法保持几何一致性；二是存在尺度幻觉，模型缺乏高空俯瞰尺度感知，易生成错误透视关系与不合理背景动态；三是长序列生成易出现内容漂移，导致首尾场景地理位置无法闭环（郑天鹏；陈雁翔；温心哲；李严成；王志远，2025），无法满足测绘对绝对位置精度的严苛要求。

0.3 主要贡献

综上，现有方法在无人机大尺度稀疏视角场景中，普遍难以兼顾三维结构准确性与生成纹理高保真度，且多数生成模型未能有效融合测绘中极易获取的物理高程数据。对此，本文提出一种由稀疏三维物理先验驱动的扩散模型视角外推方法，其核心设计理念是以离散几何锚定连续生成，即利用极少量观测影像与解算点云，在严格物理透视约束下实现高保真视角外推，主要贡献概括为以下三点：

1）**提出了融合多源物理先验的空间对齐条件构造方法。**针对传统点云渲染跨视角易出现结构拉伸与大面积空洞的问题，将机载LiDAR高精度绝对深度与全局相机外参融合构建几何空间嵌入图（GSE），再通过相对位姿变换将源光学影像重投影至目标坐标系，与GSE拼接形成空间对齐联合引导特征，在输入端实现三维物理高程与二维真实纹理的严格绑定，为扩散模型提供精准的绝对空间锚定。

2）**构建了语义与几何解耦的双路特征协同补全架构。**为解决大视角切换下遮挡空洞的高保真修复难题，设计“全局语义+局部几何”非对称注入网络：去噪阶段通过IP-Adapter提取源影像的全局材质与纹理风格先验，同时由ControlNet分支处理空间对齐特征，以残差形式注入像素级几何边界约束。该机制解耦了纹理风格迁移与几何对齐任务，使模型能在物理骨架约束下生成遮挡区域的高频纹理，保证跨视角内容连贯与透视准确。

3）**设计了基于潜空间三维透视的物理监督机制。**针对纯数据驱动模型长航线外推存在的空间漂移与尺度畸变问题，在基础噪声重建损失之外，提出潜空间几何重投影损失：利用单步去噪估计推算目标视角无噪特征，与经真实位姿变换的源视角真值特征计算掩膜误差，为扩散模型施加严格的极线几何约束，在特征层面惩罚违背LiDAR物理透视规律的生成行为，提升合成数据在下游测绘与三维重建任务中的可靠性。

1 方法模型

本章详细阐述了本文提出的基于几何坐标映射与显式约束的SDXL生成框架。该框架旨在解决无人机大尺度场景下新视角合成的几何一致性问题，通过利用稀疏视角的航拍影像，生成既符合文本描述，又在几何结构、相机位姿及语义内容上保持高度一致的新视角图像。整体方法流程包含三个核心阶段：1）几何坐标特征的构建与映射，将三维空间信息转化为二维条件；2）双流协同的条件注入机制，将几何先验显式融入预训练扩散模型；3）基于重投影一致性的优化策略，通过几何约束损失函数监督模型的训练收敛。

1.1 几何空间嵌入图

为了使预训练的SDXL模型能够感知无人机大尺度场景中的三维布局与物理尺度，本文设计了一种几何空间嵌入图（Geometric Spatial Embedding，GSE）作为主要的条件引导。GSE并非简单的图像像素集合，而是融合了视觉纹理、高精度深度以及全局空间坐标的多模态特征表示。

在无人机航拍场景中，传统的单目深度估计往往难以在数公里的空间尺度上保持绝对精度。为此，我们利用了一种结合深度学习先验与主动激光点云的融合方案。几何空间嵌入图的构造方法基于针孔相机成像几何模型设计，具体的方法流程如图1所示，总体分为四个步骤。首先，利用预训练的深度估计模型f_{DA}（基于DepthAnything网络）（Lin，Chen等，2025）获取初始相对深度图D_{raw}。其次，我们通过点云数据以及无人机的高精度位姿数据，将点云从全局世界坐标系$\boldsymbol{P}_{world}$转换至当前相机（拍摄源图像I_A时的相机位姿）的局部坐标系$\boldsymbol{P}_{cam}$。给定目标视角的相机外参，即旋转矩阵$\mathbf{R}$和平移向量$\mathbf{T}$，将点

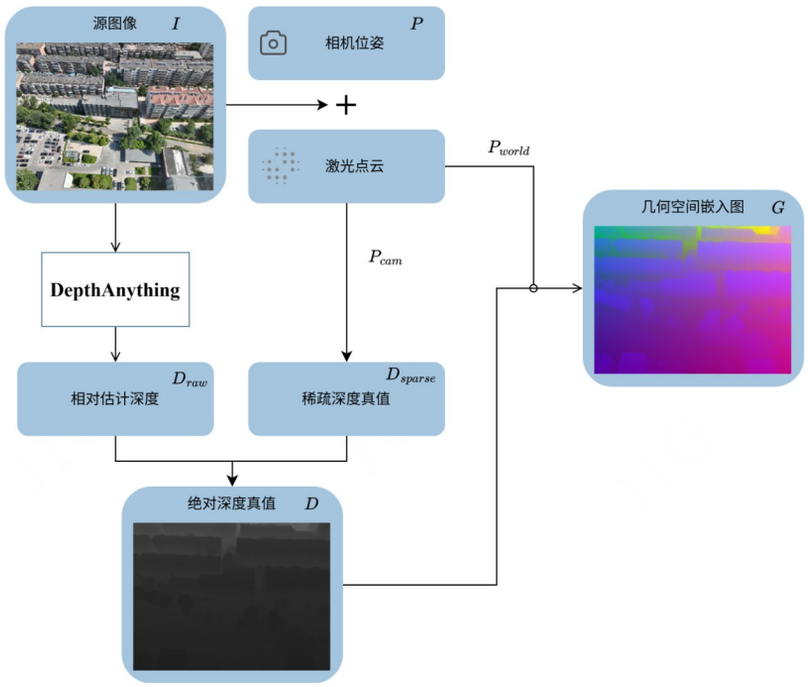


图 1　几何空间嵌入构造示意图

Fig. 1　The Illustration of the Construction of the Geometry Spatial Embedding

云从全局世界坐标系 $\boldsymbol{P}_{world}=\left[X_w, Y_w, Z_w\right]^{\mathrm{T}}$ 转换至当前相机的局部坐标系，转换公式：

$$\boldsymbol{P}_{cam}=\mathbf{R}\cdot P_{world}+\mathbf{T} \tag{1}$$

式中，$\boldsymbol{P}_{cam}=\left[x_c, y_c, z_c\right]^{T}$。只有满足 $z_c>0$（即位于相机前方）的点云会被保留。在此基础上，利用相机内参矩阵 $\boldsymbol{K}$，将三维坐标 $\boldsymbol{P}_{cam}$ 投影至二维像素平面坐标 (u, v)。投影关系满足：

$$s\begin{bmatrix}u\\v\\1\end{bmatrix}=\boldsymbol{K}\begin{bmatrix}x_c\\y_c\\z_c\end{bmatrix} \tag{2}$$

其中，s 为缩放因子，在透视投影中等于该点的相机系深度值 z_c。通过该映射，每一个落入有效图像分辨率范围（$0\leqslant u<W$, $0\leqslant v<H$）的点云点，都会在其对应的像素位置 (u, v) 上被赋予深度值 $d=z_c$。

由于激光点云具有穿透性或存在多回波现象，同一投影射线方向上可能存在多个点。我们采用Z-Buffer（深度缓冲）策略，仅保留投影到同一像素位置中 z_c 值最小，即距离相机最近的点，从而生成一张精确的稀疏深度图 D_{sparse}。由于 D_{raw}（由 DepthAnything 生成）是相对深度，而 D_{sparse} 是具有物理意义的绝对深度，两者之间存在线性的尺度和偏移关系（$d_{abs}=s\cdot d_{rel}+t$），所以我们通过利用随机采样一致性（RANSAC）算法在稀疏深度点位处进行鲁棒性拟合，剔除由动态物体或投影误差产生的噪点，计算出最优的尺度因子 s 和偏移量 t，从而获得具有物理单位的高精度绝对深度图 D。在此基础上，将绝对深度图 D 与世界坐标 $\boldsymbol{P}_{world}$ 在通道维度进行拼接由此生成的几何空间嵌入图（GSE）在空间分辨率上与输入图像严格对齐，其本质是一个多通道的几何特征张量。GSE 中的每一个像素坐标 (u, v) 不再单纯代表 RGB 颜色，而是对应着一个包含“绝对深度值 + 三维世界坐标”的高维特征向量。这一设计成功将离散的物理测绘信息密集化、张量化，使其能够作为 ControlNet 的输入条件，为后续的连续纹理生成提供几何空间锚定。

1.2　空间对齐与双路特征协同注入机制

在获取目标视角的几何骨架（GSE）后，我们需在大跨度视角切换场景下，先完成源视角局部真实纹理向目标几何骨架的精准映射，再借助扩散模型的强大生成先验，补全视场缺失、地物遮挡导致的大面积物理空洞。为此，本文提出了一种空间对齐与双路特征协同注入机制，整体框架如图 2 所示。该机制的主要目的是为二维生成模型引入严密的三维空间约束。处理流程分为两步：首先通过显式的物理空间重投影，将多源先验（纹理与三维坐标）融合为空间严格对齐的引导特征；其次采用语义与几何

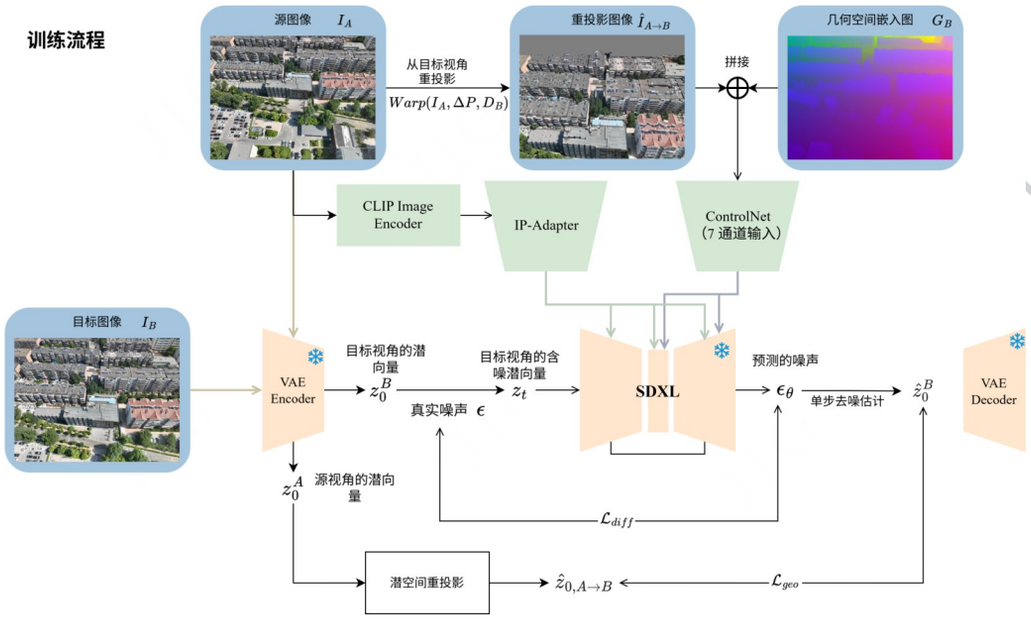


图 2　融合稀疏三维先验的视角外推方法整体架构图。

Fig. 2　Overall architecture of the generative view extrapolation method framework integrating sparse 3D priors.

解耦的双路注入架构(IP-Adapter 与 ControlNet)，引导模型在维持全局风格的同时，渲染出符合物理透视的高保真影像。

1.2.1　空间对齐的几何条件构造

在大跨度视角外推任务中，如果仅将目标视角的几何特征(G_B)单独输入网络，则模型难以在隐空间内自发建立“源视角的二维纹理”与“目标视角的三维坐标”之间的精确映射关系 。因此，在将条件特征注入扩散主干之前，本文引入了显式的物理空间对齐机制。给定源视角光学影像(I_A)、相机间的相对位姿变换矩阵($\Delta \boldsymbol{P}$)以及目标视角的绝对稠密深度图(D_B)，我们利用针孔相机的重投影映射原理，构建空间变换函数 $\mathcal{W}$。通过该函数，我们将 I_A 中的原始像素严格按照三维极线几何关系投影并对齐至目标相机的透视视锥下，得到重投影图像 $\hat{I}_{A\to B}=\mathcal{W}(I_A, \Delta \boldsymbol{P}, D_B)$。由于大基线视角切换导致的遮挡关系改变，$\hat{I}_{A\to B}$ 中不可避免地会暴露出原本被遮挡的未知区域(如图 2 重投影图像顶部灰色部分)。这些区域在重投影图像中以物理空洞的形式存在，这正是后续需要利用扩散模型强大先验进行补全的目标区域。

最后，我们将包含部分真实纹理与物理空洞的 3 通道重投影图像 $\hat{I}_{A\to B}$ 与前文构造的 4 通道目标视角下的几何空间嵌入图 G_B 在通道维度上进行拼接。由此，我们成功构建了一个空间严格对齐的 7 通道联合引导特征图。这一高维特征图在输入端，便将目标视角的几何结构与源图像的表观纹理进行了精确的空间对齐与深度融合，为后续的生成网络提供了精准的像素空间锚定，大幅降低了模型学习复杂跨视角特征映射的几何门槛。

1.2.2　基于 IP-Adapter 与 ControlNet 的双路引导前向传播

为高保真地渲染由于大跨度视角外推而暴露的物理空洞，仅依赖空间对齐的局部特征是不足的，网络还需具备对全局场景纹理风格的感知能力。为此，本文设计了“全局语义+局部几何”解耦的双路特征注入机制。

在全局语义分支，我们引入了图像提示适配器(IP-Adapter)来提取源图像 I_A 的材质与光影先验。将未经重投影的图像 I_A 输入至 CLIP 视觉特征提取器，转换为包含高维语义信息的一维图像嵌入，随后通过交叉注意力机制注入 SDXL 主干网络，为生成过程提供与参考图高度一致的全局纹理风格约束。

在局部几何分支，1.3.1 小节构造的 7 通道联合引导特征图被送入参数可训练的 ControlNet 模块。作为一个具备多尺度特征提取能力的旁路网络，ControlNet 能够精确捕获高维特征图中的像素级空间拓扑。这些多尺度几何边界特征经过零卷积(Zero-Convolutions)处理后，以残差相加的形式逐层

注入到 SDXL U-Net 的解码器模块中，从而对生成画面的物理轮廓施加严密的像素级空间控制。

在扩散模型的前向去噪传播中，目标视角的含噪潜向量 z_t^B 是主干网络的初始“画布”。在 IP-Adapter 提供的全局语义“调色盘”与 ControlNet 提供的局部几何“线稿”的协同引导下，SDXL U-Net 能够合理地推演出对应的噪声残差 ϵ_θ。这种解耦的条件注入架构，旨在引导模型在严格遵守三维物理骨架限制的同时，弥补传统图形学渲染易产生的纹理撕裂缺陷。

1.3 潜空间双重监督与联合损失函数

本文在训练阶段构建了潜空间双重监督机制，整体损失函数包含两部分。一是负责维持基础生成质量与纹理逼真度的扩散去噪重建损失；二是利用 Tweedie 公式与极线几何构建的，旨在强制约束三维空间拓扑一致性的潜空间几何重投影损失。

1.3.1 扩散去噪重建损失

在基础的潜在扩散模型（LDM）训练框架下，模型的核心优化目标是学习从含噪数据中剥离所添加的随机高斯噪声。给定目标视角的真实光学影像 I_B，首先利用冻结的 VAE 编码器将其映射至低维潜空间，获取干净的真值潜向量 z_0^B。在训练过程的任意时间步 $t \in [1, T]$，向 z_0^B 中注入从标准正态分布中采样的真实噪声 $\epsilon \sim \mathcal{N}(0, I)$，生成对应的含噪潜向量 z_t^B。

与此同时，条件提取分支分别对输入源进行特征编码：设通过 IP-Adapter 提取的全局语义嵌入特征为 c_{sem}，通过 ControlNet 处理 7 通道联合引导图所提取的局部空间特征为 c_{geo}。

随后，将 z_t^B 作为初始“画布”，连同时间步 t 和双路条件（c_{sem}, c_{geo}）共同送入 SDXL 主干网络 ϵ_θ 中。络经过前向传播输出当前步的预测噪声，并通过计算该预测噪声与真实噪声 ϵ 之间的均方误差（MSE）来构建基础去噪损失 $\mathcal{L}_{diff}$：

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0^B, \epsilon, t, c_{sem}, c_{geo}}\left[\left\| \epsilon - \epsilon_\theta\left(z_t^B, t, c_{sem}, c_{geo}\right)\right\|_2^2\right] \tag{3}$$

该损失函数提供了模型训练的基础监督信号，主要负责驱动网络学习复杂的高频纹理分布规律，确保在双模态条件的约束下，生成的内容能够最大限度地还原目标视角下的真实地物外观与光影细节。

1.3.2 潜空间几何重投影损失

仅依赖 $\mathcal{L}_{diff}$ 虽然能够保证单视角的图像质量，但无法严格约束源视角 A 与目标视角 B 之间的物理几何对应关系。为此，本文在潜在空间引入了显式的重投影约束。由于直接对含噪状态 z_t^B 进行几何变换缺乏物理意义，我们首先利用 Tweedie 公式，根据当前网络预测的噪声 ϵ_θ，在时间步 t 估算出目标视角无噪的预测潜向量 $\hat{z}_0^B$：

$$\hat{z}_0^B = \frac{z_t^B - \sqrt{1 - \overline{\alpha}_t}\,\epsilon_\theta}{\sqrt{\overline{\alpha}_t}} \tag{4}$$

随后，将源视角的真实潜向量 z_0^A 作为参考基准。如前文提到的针孔相机的重投影映射原理，利用相机间的相对位姿变换矩阵 $\Delta \boldsymbol{P}$ 以及目标视角的深度图 D_B，构建潜空间变换函数 $\mathcal{W}_{\text{latent}}$，将源视角的真实特征重投影至目标视角的坐标系下，得到变换后的潜向量 $\hat{z}_{0,A\to B}$。

为了使得模型生成的几何结构与真实物理变换保持一致，我们计算预测潜向量 $\hat{z}_{0,A\to B}$ 与变换潜向量 $\hat{z}_0^B$ 之间的掩膜均方误差，定义为几何重投影损失 $\mathcal{L}_{geo}$：

$$\mathcal{L}_{geo} = \left\| \mathcal{M} \odot \left(\hat{z}_0^B - \hat{z}_{0,A\to B}\right)\right\|_2^2 \tag{5}$$

式中，$\odot$ 表示哈达玛积。$\mathcal{M}$ 为可视性掩膜，用于排除视角切换导致的遮挡区域或超出视场的部分，确保损失计算仅在两个视角共同可见的共视区域内进行。

最后，本方法的总损失函数 $\mathcal{L}$ 由上述两部分加权组合而成：

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda \mathcal{L}_{geo} \tag{6}$$

式中，λ 为平衡两个优化目标的超参数，$\mathcal{L}_{diff}$ 负责维持模型强大的高保真纹理生成能力，而 $\mathcal{L}_{geo}$ 则作为物理正则化项，迫使生成的特征在三维空间中严格贴合相机位姿先验，从而实现高精度的多视角一致性。

1.4 推理阶段

与训练阶段利用目标图像加噪不同，推理阶段在仅依靠先验条件，从纯高斯噪声中渲染（估计）出高质量的目标视角影像。推理阶段的整体输入包含三个部分——源视角的单张参考光学图像 I_A，目标视角的相机位姿 P_{cam}^B，以及场景的点云先验。

如图 3 所示，整体流程分为三个步骤。首先，参照 2.1 章节的方法，我们先从输入的源图像 I_A 得到

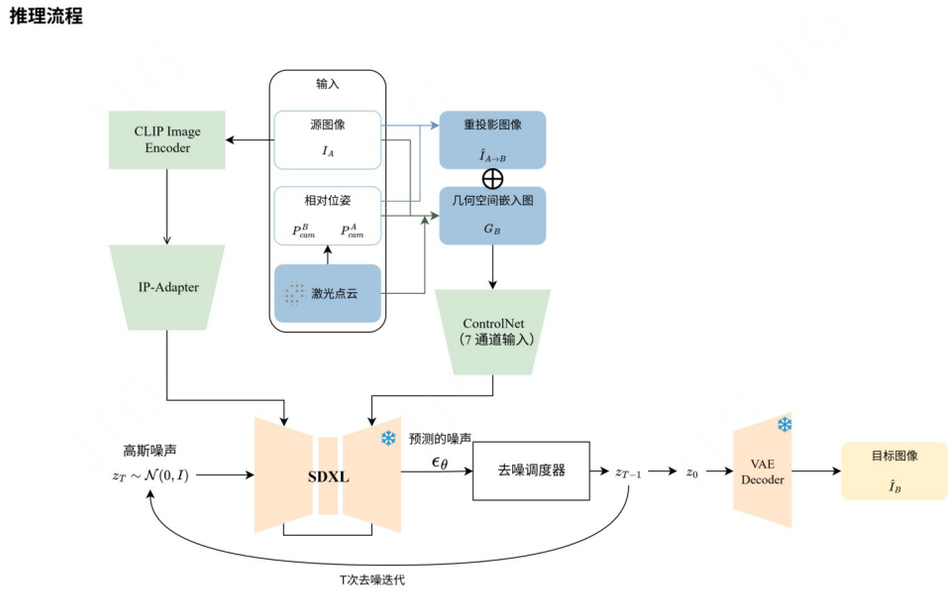


图 3　方法模型的推理流程图示

Fig. 3　The illustration of the inference phase of the method workflow

对应的绝对深度图 D_A，再从 $\boldsymbol{P}^B_{cam}$ 中解算得到稀疏深度真值 D^B_{sparse}。随后，利用相对位姿 $\Delta\boldsymbol{P}$（A 视角和 B 视角的相对位姿）将绝对深度 D_A 正向重投影至目标视角获得该视角下的深度 $\hat{D}_{A\to B}$。然后进一步将 $\hat{D}_{A\to B}$ 和 D^B_{sparse} 融合得到 D_B，最后与世界坐标拼接为目标视角 B 的几何空间嵌入图 G_B。

输入处理之后，就进入到了双路特征协同注入环节。如图 2 上半部分所示流程，我们通过参考源图像 I_A 和目标的几何空间嵌入图 G_B 获取双路条件特征 (c_{sem}, c_{geo})，它们分别通过 IP-Adapter 和 ControlNet 注入到 SDXL 主干网络中。在历经 T 步去噪迭代，预测并剥离噪声后，最终获得无噪声潜向量 z_0。最后 z_0 再通过 SDXL 中 VAE 解码器映射回到 RGB 空间，即输出最终的目标视角图像 $\hat{I}_B$。

2　实验结果

2.1　实验实现细节

为了本实验的可复现性，本章节将从数据集、实验环境、数据预处理、实验评估指标，以及方法实现细节 5 个方面详细地描述本方法实验是如何进行的。

2.1.1　数据集与实验环境

为了验证本文提出方法在真实复杂场景下的有效性与鲁棒性，我们构建了一个大规模、多地貌特征的无人机倾斜摄影数据集。

实验数据采集区域位于中国山东省济南市，覆盖总面积约 20 平方公里，分为 8 个片区，共计 80405 张光学影像。该区域具有显著的地物异质性，其中约 15 平方公里为建筑密集、路网复杂的城市核心区，包含高层写字楼、居民住宅及交通枢纽；另约 5 平方公里为包含农田、植被及低矮建筑的城郊村落区域。这种“城市-乡村”混合的场景构成，确保了数据集涵盖了丰富的纹理特征与多样的拓扑结构，能够充分测试模型对不同尺度地物的泛化能力。数据采集设备采用大疆 M3E 无人机平台，获取了高分辨率光学航拍影像及同步的定位定姿系统数据。同时，为了获取精确的几何真值，我们获取了该区域相应的高精度三维点云数据。在预处理阶段，原始光学影像保持原始分辨率与色彩配置，点云数据则利用 DJI Terra 软件进行解算与配准，并统一转换为标准“. las”格式，作为后续生成几何空间嵌入图的基准数据源。我们将数据集按照 9∶1 的比例随机划分为训练集与测试集，确保评估结果的客观性。

所有实验均在基于 Linux Ubuntu 系统的深度学习服务器上进行，硬件核心配备四张张 NVIDIA GeForce RTX 4090 GPU（24GB 显存）。软件环境基于 Python3. 8 与 PyTorch 深度学习框架，并利用 Diffusers 库构建潜在扩散模型管线。在训练策略上，得

益于本文采用的ControlNet旁路注入架构，我们无需对庞大的SDXL主干网络（约26亿参数）进行全量微调，仅需更新ControlNet分支及部分投影层的参数。这显著降低了显存开销与计算成本，使得在单卡环境下进行高分辨率训练成为可能。实验采用AdamW优化器，学习率设定为1×10^{-5}，并配合梯度累积策略以稳定训练过程。模型最终训练迭代步数设置为30，000步，此时损失函数趋于收敛，生成的几何结构与纹理质量达到最佳平衡。

2.1.2　数据预处理

首先，针对无人机拍摄的原始高分辨率航拍影像（原始分辨率为5280 × 3956），为了平衡计算负载并适配显存限制，我们对其进行了固定比例的降采样处理。相应地，相机的内参矩阵$\boldsymbol{K}$也进行了同比例的缩放更新，以确保像素坐标系与图像分辨率的对应关系；而相机的外参矩阵（$\mathbf{R}, \mathbf{T}$）作为表征相机在世界坐标系下物理位置与姿态的参数，则保持不变。这一处理确保了后续输入的相机位姿与缩放后的二维影像在几何投影关系上严格对齐。接下来，为了构建具备物理尺度的稠密深度图，如1.2.1所述本文提出了一种“相对-绝对”深度融合策略，构建了一种几何空间嵌入图，该嵌入图编码了像素级的三维空间关系，作为几何特征提取的重要依据。

2.1.3　实验评估指标

为了全面评估生成图像的视觉质量、几何准确性以及多视角一致性，本文采用了主观与客观相结合的评估体系。我们将测试集中的相机位姿与文本提示输入到本方法及对比模型中，生成相应的多视角图像序列，随后通过以下三个维度的指标进行定量分析：

其一是重建保真度与几何一致性评估，为了客观衡量生成图像与真实地理场景的几何对应关系，我们在测试集的保留视角上进行像素级与特征级的全参考评估。具体而言，利用给定的相机位姿生成目标视图，并将其与真实拍摄的Ground Truth（GT）图像进行对比。我们计算三个核心指标：学习感知图像块相似度（LPIPS）、峰值信噪比（PSNR）以及SSIM（结构相似性）。LPIPS利用预训练的VGG网络提取深层特征并计算距离。相比于像素误差，LPIPS对图像的几何畸变与结构偏移高度敏感。LPIPS值越低，直接证明模型生成的新视角在几何结构与纹理布局上与真实场景越吻合。PSNR与SSIM分别衡量生成图像与真值在像素强度的对齐程度以及结构信息的完整性。

其二是基于视觉语言模型的感知质量评估（LLaVA-IQA）。针对生成图像的纹理细节与视觉逼真度，我们采用先进的视觉语言模型（VLM）进行无参考质量评估。参考VistaDream等近期工作（Wang，Liu等，2025），我们选取LLaVA模型作（Liu，Li等，2023）为“评审员”，从五个维度——无噪声程度、边缘锐度、结构合理性、细节丰富度，以及图像质量，对重建场景的渲染视图进行评分。具体操作如下：首先，从上面提到的方法得到稀疏点云，再重建为三维场景，在每个场景沿着同样的预定轨迹均匀采样50个新视角进行渲染。将渲染图像输入LLaVA，并设计如下提示词模板：“图像名”加“问题”，然后加请用是或否来回答这个问题。其中“问题”对应五个评估维度：噪声、边缘锐度、结构合理性、细节丰富度、以及综合质量。比如询问噪声程度时，则问：图像A是否没有明显的噪声或伪影失真？最终，我们计算模型回答“是”的比例作为该维度的量化得分。

最后是语义一致性指标。为了验证生成图像是否准确响应了文本提示（如地理环境描述），我们采用经典的CLIP（ViT-L/14）模型（Radford，Kim等，2021）进行语义对齐测试。具体的操作是通过计算生成图像的视觉嵌入与输入文本提示的文本嵌入之间的余弦相似度，来表明生成内容与输入语义的符合程度，如下所表述：

$$\text{CLIP} - \text{Score} = \frac{e_{img} \cdot e_{txt}}{\left\| e_{img} \right\|_2 \cdot \left\| e_{txt} \right\|_2} \tag{7}$$

式中，e_{img}表示由CLIP模型的图像编码器提取出的生成图像的视觉特征向量；e_{txt}则是由文本编码器提取出的输入提示词的文本特征向量。该指标越高，表明生成内容在语义上越符合用户的文本描述，且未出现与场景描述无关的异常物体。

2.2　定性实验

本章节我们用两个定性实验来验证本文方法在稀疏大基线视角下的实际生成效果与实际可用性。第一个实验主要是直观对比各方法生成的画面够不够逼真、场景的几何结构有没有变形；第二个实验则是把生成的图像直接给到下游的三维重建算法，通过观察最终建出来的点云和视锥范围，来看看它是

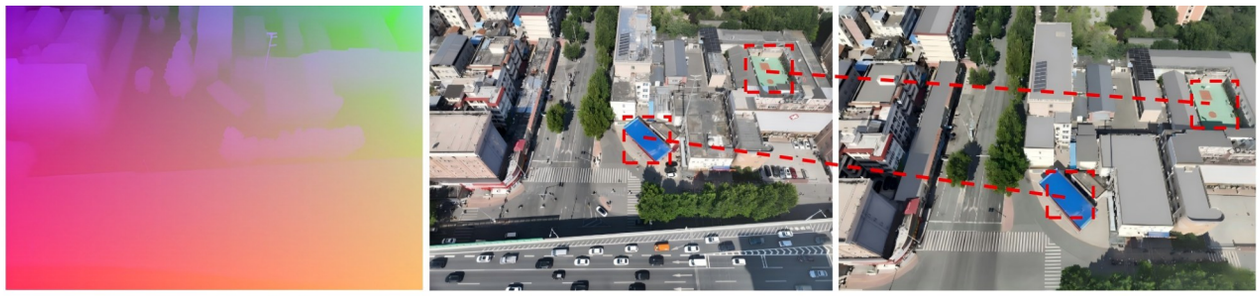
否能真正帮我们把场景的三维空间给拓展出去。

2.2.1　跨视角生成方法的定性评估与结构保持分析

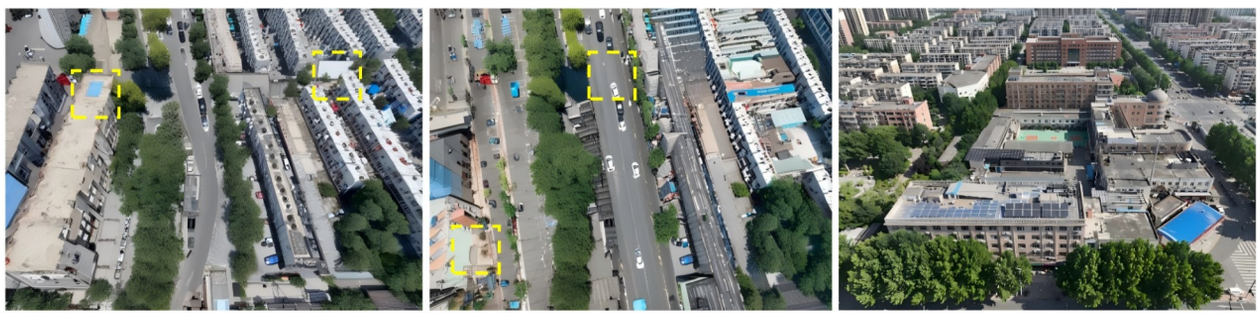
图4展示了在给定单张参考图像、相机位姿变化以及一个基于场景点云的几何先验的情况下，四种方法外推的图像结果。我们的方法和SDXL+ControlNet作为可以处理展示了在给定参考视角图像、目标相机位姿变化及基于场景点云的几何先验条件下，不同方法生成图像的对比结果。我们从测试集中选取了3个具有代表性的场景。以第一个场景为例，我们给这4种方法输入了参考源图像以及对应的相机位姿变化，输入的偏移量是向经度方向移动40米——即目标视角相当于在保持当前相机姿态不变的情况下，相机往经度方向平移了40米。对于VistaDream与基线方法SDXL，由于其网络架构本身不具备融合三维几何模态的输入条件，生成过程只能依赖参考源图像与文本相机位姿提示。同时，我们构建了SDXL + ControlNet这个基线方法，并为其提供了与本文方法完全对等的多模态输入——即在图像与位姿之外，同步注入了由场景点云解算得到的目标视角几何深度图。需要特别说明的是，在本次所有测试中，设定的目标相机位姿，对应的都是真实无人机航线中参考图往后的第二帧的拍摄位置。也就是像前面第一个场景描述的那样：相机姿态基本保持不变，仅仅是位置在一个方向上向前平移。

通过对比结果我们可以看到，参考视角中的红框地物元素在我们的合成结果中都得到了保留，并且从整体上看我们的合成结果的观察视角和输入的相机位姿偏移基本符合，即仅仅是在观察纵深上往前移动一段距离，最后表现为红框中的地物在画面中的位置往后倒退了一点。相比之下，我们可以观察到VistaDream方法生成的图像从内容上、色彩上，虽能和参考图像保持一致，但其相机的位姿明显和目标视角不同。从基线方法SDXL的生成结果可以看到生成的图像在色调风格上有一定的相似性，然而图像中呈现了各种几何错误。比如第一、五个场景中的黄框中路与建筑交织在一起，第二个场景的黄框中路面发生了扭曲以及第三个场景的黄框路面断开并在另一面续上了。这些情况都是在缺乏有效的几何与语义约束发生的“漂移”现象。而“SXDL+ControlNet”加入了与本方法同样的几何先验，但是没有加入几何约束，所以还是暴露了显著的几何错误。

上述对比结果清楚地表明，在大基线视角外推这种极端任务中，简单地向模型堆砌三维先验数据并不能从根本上解决问题。对于二维扩散模型而言，仅仅“看到”深度图也是不够的，本方法在底层架构上施加严格的物理透视约束，解决了生成网络在陌生视角下产生的结构畸变，合成经得起几何推敲的画面。

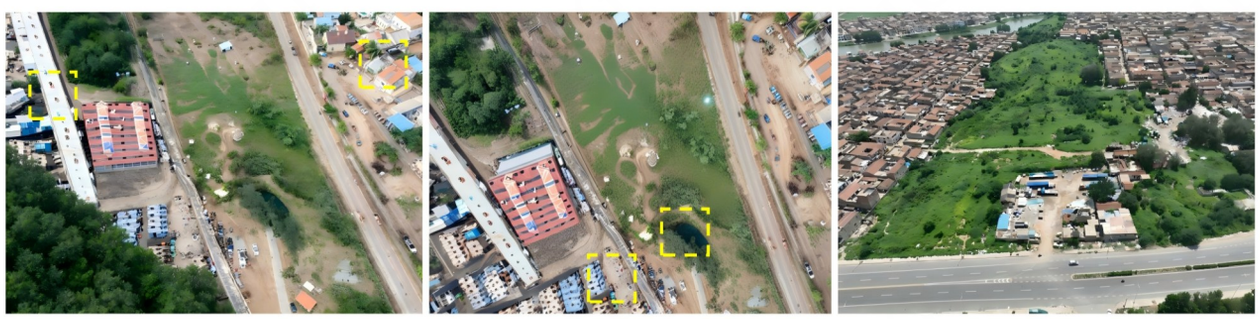


几何先验　　输入视角　　我们的方法



SDXL+ControlNet　　SDXL　　VistaDream

几何先验　　输入视角　　我们的方法



SDXL+ControlNet　　SDXL　　VistaDream

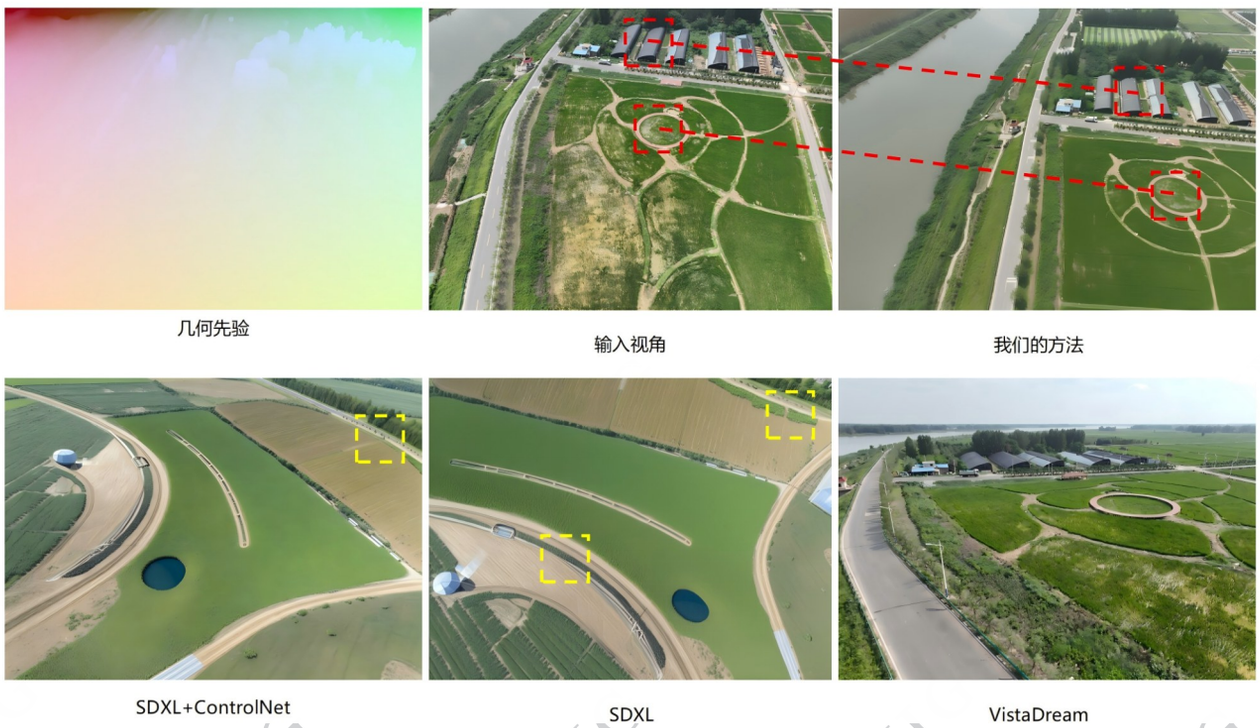


图 4　给定单张参考图像与相机位姿变化信息，本实验对比了四种生成方法的结果。

Fig. 4　Given a single reference image and camera pose variation, this experiment compares the synthesized results of four generation methods.

2.2.2　基于 VGGT 视锥分析的场景空间拓展能力评估

为了进一步验证本框架生成的新视角图像是否具备严格的三维多视图几何一致性，以及其在实际航测工程中是否能够真正辅助下游的三维任务，本文设计了一组基于 VGGT 重建框架（Wang, Chen 等，2025）的场景空间拓展定性实验。在该实验中，我们采用 VGGT 作为统一的评估工具。其核心逻辑在于：利用输入图像进行稀疏特征提取与重建，并通过观察最终输出的三维点云覆盖范围以及估计的相机视锥（Camera Frustum）分布，来反推输入图像序列的三维空间信息量与几何合理性。为了形成严谨的对照，我们针对同一真实场景构建了三种不同的输入条件，其一是理想对照组（如图 5 a 所示），输入 4 张真实影像，其二是极端稀疏基线组（如图 5 b 所示），仅输入 1 张真实影像，最后则是实验组（如图 5 c 所示），输入 1 张真实影像加上 3 张本方法合成的影像。

首先我们从图 5 a 可以看到，作为该场景的三维重建上限，该组直观展示了在拥有充足多视角真实观测时，VGGT 所能恢复的完整点云范围与真实的相机视锥分布。然后图 5 b 作为重建下限该组展示当数据采集极度受限（仅有单视点覆盖）时，通过重建算法得到的三维场景跟上一组比起来，如红色虚线框标注的，丢失了大块可见区域，毕竟重建模型没有“看见”这些地方的图像，所以也没法无中生有。最后在本方法参与的重建实验组中，我们将单张真实影像与本框架生成的外推图像混合作为输入，图 5 c 展示的重建结果接近理想对照组的三维重建范围与视锥排列。这足以验证在仅拥有单视角真实观测的极端条件下，本文生成的“虚拟多视角图像”能够充当有效的几何锚点，从而辅助 VGGT 算法重建出更大范围的三维场景。

2.3　定量实验

2.3.1　重建保真度与几何一致性评估

为了客观评估生成图像的质量，我们在测试集上计算了 PSNR、SSIM 以及 LPIPS 三项核心评价指标，对比结果如表 1 所示。需要特别说明的是，在大跨度视角外推这一极端任务中，由于存在大面积的物理遮挡与未知区域，模型必须进行大量的“生成式”纹理推演，而非简单的多视图像素插值。这种生成范式虽然能合成逼真的高频细节，但由于生成的

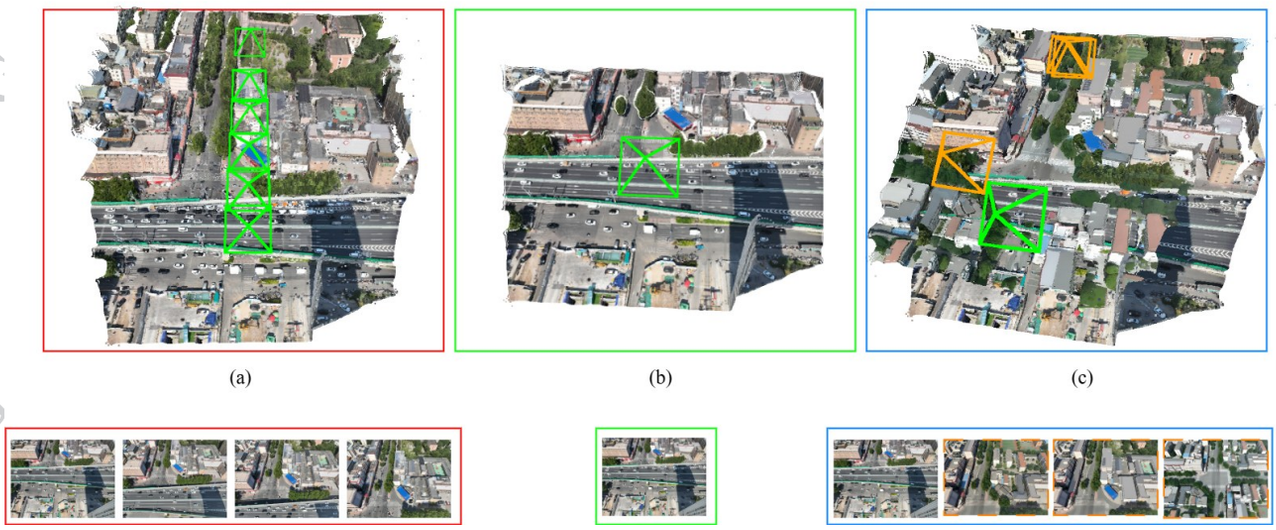


图 5　基于 VGGT 的场景三维重建空间拓展评估。

Fig. 5　Spatial expansion evaluation of 3D scene reconstruction based on VGGT.

纹理(如具体的树叶分布、建筑表面材质)在微观位置上难以与真实图像达成严苛的像素级绝对对齐,因此包括本文方法在内的所有基于扩散模型的方法,在传统的全参考像素级惩罚指标(如 PSNR)上的绝对数值均不突出。

尽管面临这一苛刻的指标特性,在横向对比中,本方法依然展现出了显著的优越性。在衡量结构保真度的 SSIM 指标上,本文方法达到了 0.1918,相较于基础 SDXL(0.0743)和主流多视角生成方法 VistaDream(0.1376)分别大幅提升了约 158% 和 39%。这表明,纯数据驱动的 SDXL 因缺乏物理骨架约束,极易产生导致结构错位的空间“幻觉”;而 VistaDream 虽然引入了体素特征,但受限于体素网格的分辨率上限,难以在大尺度无人机航拍场景中锐利地保持精细地物的边缘。相比之下,本文注入的几何空间嵌入图为去噪过程提供了精确的局部坐标锚定,有效遏制了空间漂移。更重要的是,在能够更好反映人类视觉感知质量与特征相似度的 LPIPS 指标上,本方法取得了 0.4658 的最优成绩。LPIPS 误差的显著降低有力地证明了,摒弃了传统像素级死板对齐的束缚后,本文框架渲染出的画面在全局语义布局与局部高频纹理上,都具备着极高的感知保真度,最接近真实拍摄的航拍视觉体验。

2.3.2　基于视觉语言模型的感知质量评估

如表2所示,与仅依赖二维图像特征的SDXL及SDXL+ControlNet相比,本文方法在结构合理性上有较大的优势(0.611vs0.406/0.483)。这证明了SDXL在添加ControlNet模块的情况下,在面对大尺度无人机航拍影像时,仅靠单视角和深度图难以维持复杂的路网和建筑拓扑。本文方法之所能取得改进,核心在于跨视角几何引导以及双优化目标的设计。通过强制对双视角在潜空间进行三维几何变换与特征对齐,模型在生成时不仅顾及了单帧的逼真度,更掌握了地物的物理透视关系,从而生成了轮廓更锐利、空间布局更符合常理的场景。相比于先进的VistaDream,本文方法在无噪声程度(0.851vs0.889)与细节丰富度(0.890vs0.927)上略逊一筹。这在生成式模型中是一个常见的“权衡”现象:VistaDream倾向于生成极具视觉冲击力但可能缺乏严格几何约束的平滑纹理,从而在纯二维的细节和无噪点评估中获得高分,但是其对于相机的位姿的控制还需进一步优化。

表1　各方法在几何一致性LPIPS / PSNR / SSIM指标上的定量评估结果

Table 1　Quantitative evaluation results of each method on geometric consistency metrics: LPIPS / PSNR / SSIM

方法	PSNR(↑)	SSIM(↑)	LPIPS(↓)
VistaDream	11.524	0.1376	0.6886
sdxl	9.6415	0.0743	0.7249
sdxl+controlnet	9.7315	0.0809	0.7149
我们的方法	13.125	0.1918	0.4658

表2　各方法在LLaVA-IQA五维感知质量指标上的定量评估结果

Table 2　Quantitative evaluation results of each method on the five-dimensional perceptual quality metrics of LLaVA-IQA.

方法	无噪声程度	边缘锐度	结构合理性	细节丰富度	图像质量
VistaDream	0.889	0.315	0.576	0.927	0.613
sdxl	0.792	0.208	0.406	0.825	0.531
sdxl+controlnet	0.848	0.256	0.483	0.832	0.583
我们的方法	0.851	0.342	0.611	0.89	0.642

2.3.3　语义一致性CLIP-Score

如表3所示本文方法在CLIP-Score指标上取得了0.2860的最高分，优于基线模型SDXL(0.2229)及当前先进的VistaDream方法(0.2600)，证明了本方法在保持高精度几何对齐的同时，不仅没有削弱基础扩散模型的文本理解能力，反而显著增强了图像的语义表现力。

表3　各方法在语义一致性(CLIP-Score)上的定量评估结果

Table 3　Quantitative evaluation results of each method on semantic consistency (CLIP-Score).

	VistaDream	sdxl	sdxl+controlnet	我们的方法
CLIP-Score	0.2600	0.2229	0.2411	0.286

2.4　消融实验

根据章节1的方法部分，我们将本文的完整方法与三种去除了关键组件的变体模型进行了对比：去除点云绝对尺度校正的变体(w/o LiDAR)、去除世界坐标嵌入的变体(w/o Coord)以及去除重投影一致性损失的变体(w/o $\mathcal{L}_{geo}$)。定量对比结果如表4所示，本文的完整方法在反映图像像素精度的PSNR、反映结构完整性的SSIM、反映感知真实度的LPIPS，以及反映语义对齐的CLIP-Score上，均取得了最高分。消融结果验证了各模块在框架中的不可替代性：首先，移除世界坐标嵌入(w/o Coord)使PSNR骤降至7.676，表明缺乏全局锚定极易导致地物无法准确映射，引发严重的像素级空间漂移；其次，去除激光点云绝对尺度校正(w/o LiDAR)导致SSIM(0.114)与CLIP-Score(0.122)下跌，印证了单纯依赖相对深度会产生强烈的尺度歧义，进而引发严重的几何结构畸变与语义识别错误；最后，去除重投影一致性损失(w/o $\mathcal{L}_{geo}$)令LPIPS恶化至最差的0.633，证明缺乏跨视角的三维物理监督会使模型退化为孤立的二维“盲猜”，打破多视角间的纹理连续性并极易产生幻觉内容，大幅降低图像的整体感知质量。

表4　核心模块消融实验的定量评估结果

Table 4　Quantitative evaluation results of the ablation experiments on the core modules.

方法状态	PSNR(↑)	SSIM(↑)	LPIPS(↓)	CLIP-Score(↑)
w/o LiDAR	8.746	0.114	0.585	0.122
w/o Coord	7.676	0.157	0.596	0.172
w/o $\mathcal{L}_{geo}$	10.001	0.139	0.633	0.135
完整方法	13.125	0.192	0.466	0.286

3　结　论

在大规模无人机实景三维重建与测绘任务中，受限于数据采集成本与航线规划，获取具备超高重叠率的密集影像往往面临极大挑战。针对利用稀疏观测进行大基线视角外推时面临的图形学纹理撕裂与二维生成空间漂移问题，本文提出了一种融合稀疏三维物理先验的生成式视角外推方法。本研究突破了传统纯数据驱动生成的局限，成功验证了“以离散物理骨架引导连续像素生成”的技术可行性。本方法构建了精准的多源空间对齐机制，实现了语义与几何解耦的双路特征协同以及确立了潜空间双重监督的训练范式。在基于真实无人机航拍数据集的定性与定量评估中，本文方法在处理大跨度视点外推时，不仅在全参考指标(PSNR，SSIM，LPIPS)上显著优于现有基线模型，更在视觉感知与语义一致性上达到了较高的程度。综上所述，本研究为无人机稀疏视角下的高保真多视图合成提供了一种物理自洽的新范式，在降低实景三维密集数据采集成本、提升复杂场景渲染质量方面具有重要的工程应用价值与广阔的发展前景。

尽管本文方法在几何一致性受控生成上取得了实质性进展，但在实际应用中仍存在局限，这也是我们未来工作的重点研究方向。首先是长序列自回归

生成的误差累积,本文的对偶视角约束能够有效保证参考帧与相邻下一帧之间的高度一致性。然而,在进行长航线推演时,若持续以模型合成的结果作为下一帧的参考输入(自回归迭代),微小的几何与纹理误差会不断累积,最终导致生成画面与真实物理场景发生逐渐偏离。有趣的是,这种“发散”特性也为创造具有相似地貌特征但拓扑结构全新的虚拟场景提供了潜力。未来我们将探索引入全局场景表征来锚定长序列生成的绝对边界。第二点是微观纹理的逼真度瓶颈,受限于潜扩散模型的压缩特性与模型固有的平滑倾向,尽管生成的图像在宏观结构上准确,但在微观细节(如植被叶片纹理、建筑外立面材质、细小电缆等)上仍带有一定的“生成式涂抹感”或伪影,难以达到完全媲美真实摄影的照片级逼真度。最后为对动态场景的静态假设局限,本方法的重投影一致性损失严格建立在“场景绝对静态”的刚体变换假设之上。当无人机拍摄的数据中包含运动的车辆或行人时,这些动态地物会破坏极线几何约束,导致模型在生成时出现重影或运动模糊。如何在复杂的特征对齐中实现“动静解耦”,过滤或单独处理动态物体,是进一步提升模型鲁棒性的关键挑战。

参考文献(References)

Bernal-Berdun E, A Serrano, B Masia, M Gadelha, Y Hold-Geoffroy, X Sun, et al. Year. PreciseCam: Precise Camera Control for Text-to-Image Generation//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 2724-2733 [10.1109/CVPR52734.2025.00260: 10.1109/CVPR52734.2025.00260]

Blattmann A, T Dockhorn, S Kulal, D Mendelevitch, M Kilian, D Lorenz, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127:

Chan E R, C Z Lin, M A Chan, K Nagano, B Pan, S d Mello, et al. Year. Efficient Geometry-aware 3D Generative Adversarial Networks//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 16102-16112 [10.1109/CVPR52688.2022.01565: 10.1109/CVPR52688.2022.01565]

Chen Z, Y Wang, F Wang, Z Wang, F Sun and H Liu. 2025. V3D: Video Diffusion Models are Effective 3D Generators. IEEE Transactions on Pattern Analysis and Machine Intelligence: 1-18 [10.1109/TPAMI.2025.3581312: 10.1109/TPAMI.2025.3581312]

Deitke M, D Schwenk, J Salvador, L Weihs, O Michel, E VanderBilt, et al. Year. Objaverse: A Universe of Annotated 3D Objects//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 13142-13153 [10.1109/CVPR52729.2023.01263: 10.1109/CVPR52729.2023.01263]

Hong Y, K Zhang, J Gu, S Bi, Y Zhou, D Liu, et al. 2023. Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400:

Jiang Q Y, Y He, G Li, J Lin, L Li and W J Li. Year. SVD: A Large-Scale Short Video Dataset for Near-Duplicate Video Retrieval//2019 IEEE/CVF International Conference on Computer Vision (ICCV). City: 5280-5288 [10.1109/ICCV.2019.00538: 10.1109/ICCV.2019.00538]

Kerbl B, G Kopanas, T Leimkuehler and G Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph., 42 (4): Article 139 [10.1145/3592433: 10.1145/3592433]

Li X, Q Zhang, D Kang, W Cheng, Y Gao, J Zhang, et al. 2024. Advances in 3d generation: A survey. arXiv preprint arXiv: 2401.17807:

Lin C H, J Gao, L Tang, T Takikawa, X Zeng, X Huang, et al. Year. Magic3D: High-Resolution Text-to-3D Content Creation//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 300-309 [10.1109/CVPR52729.2023.00037: 10.1109/CVPR52729.2023.00037]

Lin H, S Chen, J Liew, D Y Chen, Z Li, G Shi, et al. 2025. Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647:

Liu H, C Li, Q Wu and Y J Lee (2023). Visual Instruction Tuning. Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New Orleans, Louisiana, USA, Neural Information Processing Systems Foundation, Inc. (NeurIPS). 36: 34892-34916.

Liu R, R Wu, B V Hoorick, P Tokmakov, S Zakharov and C Vondrick. Year. Zero-1-to-3: Zero-shot One Image to 3D Object//2023 IEEE/CVF International Conference on Computer Vision (ICCV). City: 9264-9275 [10.1109/ICCV51070.2023.00853: 10.1109/ICCV51070.2023.00853]

Liu Y, C Lin, Z Zeng, X Long, L Liu, T Komura, et al. Year. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image//ICLR. City:

Long X, Y C Guo, C Lin, Y Liu, Z Dou, L Liu, et al. Year. Wonder3D: Single Image to 3D Using Cross-Domain Diffusion//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 9970-9980 [10.1109/CVPR52733.2024.00951: 10.1109/CVPR52733.2024.00951]

Mildenhall B, P P Srinivasan, M Tancik, J T Barron, R Ramamoorthi and R Ng. Year. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis//City: Springer International Publishing: 405-421 [10.1007/978-3-030-58452-8_24: 10.1007/978-3-030-58452-8_24]

Podell D, Z English, K Lacey, A Blattmann, T Dockhorn, J Müller, et

al. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952:

Poole B, A Jain, J T Barron and B Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988:

Radford A, J W Kim, C Hallacy, A Ramesh, G Goh, S Agarwal, et al. Year. Learning transferable visual models from natural language supervision//International conference on machine learning. City: PmLR: 8748-8763

Rombach R, A Blattmann, D Lorenz, P Esser and B Ommer. Year. High-Resolution Image Synthesis with Latent Diffusion Models// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 10674-10685 [10.1109/CVPR52688.2022.01042: 10.1109/CVPR52688.2022.01042]

Shen T, J Gao, K Yin, M-Y Liu and S Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in neural information processing systems, 34: 6087-6101

Shi Y, P Wang, J Ye, M Long, K Li and X Yang. 2023. Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512:

Tang J, Z Chen, X Chen, T Wang, G Zeng and Z Liu. Year. LGM: Large Multi-view Gaussian Model for High-Resolution 3D Content Creation//Computer Vision – ECCV 2024. City: Springer Nature Switzerland: 1-18 [10.1007/978-3-031-73235-5_1: 10.1007/978-3-031-73235-5_1]

Voleti V, C-H Yao, M Boss, A Letts, D Pankratz, D Tochilkin, et al. Year. SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image Using Latent Video Diffusion//Computer Vision – ECCV 2024. City: Springer Nature Switzerland: 439-457 [10.1007/978-3-031-73232-4_25: 10.1007/978-3-031-73232-4_25]

Wang H, Y Liu, Z Liu, W Wang, Z Dong and B Yang. Year. Vistadream: Sampling multiview consistent images for single-view scene reconstruction//Proceedings of the IEEE/CVF International Conference on Computer Vision. City: 26772-26782

Wang J, M Chen, N Karaev, A Vedaldi, C Rupprecht and D Novotny. Year. VGGT: Visual Geometry Grounded Transformer//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 5294-5306 [10.1109/CVPR52734.2025.00499: 10.1109/CVPR52734.2025.00499]

Wang Z, C Lu, Y Wang, F Bao, C Li, H Su, et al. (2023). ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New Orleans, Louisiana, USA, Neural Information Processing Systems Foundation, Inc. (NeurIPS). 36: 8406-8441.

Xu J, W Cheng, Y Gao, X Wang, S Gao and Y Shan. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191:

Yang J, Z Cheng, Y Duan, P Ji and H Li. Year. ConsistNet: Enforcing 3D Consistency for Multi-View Images Diffusion//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). City: 7079-7088 [10.1109/CVPR52733.2024.00676: 10.1109/CVPR52733.2024.00676]

Ye J, P Wang, K Li, Y Shi and H Wang. Year. Consistent-1-to-3: Consistent Image to 3D View Synthesis via Geometry-aware Diffusion Models//2024 International Conference on 3D Vision (3DV). City: 664-674 [10.1109/3DV62453.2024.00027: 10.1109/3DV62453.2024.00027]

Huang Y, Guo Y, Lu Y, Jiang P, Wang F. 2025. 3D reconstruction of neural radiation fields constrained by the Manhattan structure in urban remote sensing images. Journal of Image and Graphics, 30(7): 2584-2596 (黄洋, 郭宇, 路遥, 姜鹏, 王飞. 2025. Manhattan结构约束神经辐射场在城市遥感图像中的三维重建. 中国图象图形学报, 30(7): 2584-2596) [DOI: 10.11834/jig.240544]

Zhan R Y, Fan Y, Zhou L N, Xie Y B, Chen J X, Yang H Y, et al. 2026. Dynamically distribution-aware quantization for diffusion models. Journal of Image and Graphics, 31(3): 0745-0754 (占瑞乙, 樊轶, 周丽娜, 谢宇宝, 陈佳鑫, 杨鸿宇, 等. 2026. 分布范围动态感知的扩散模型量化. 中国图象图形学报, 31(3): 0745-0754) [DOI: 10.11834/jig.250319]

Zheng T P, Chen Y X, Wen X Z, Li Y C, Wang Z Y. 2025. Diffusion model-generated video dataset and detection benchmarks. Journal of Image and Graphics, 30(4): 1059-1071 (郑天鹏, 陈雁翔, 温心哲, 李严成, 王志远. 2025. 扩散模型生成视频数据集及其检测基准研究. 中国图象图形学报, 30(4): 1059-1071) [DOI: 10.11834/jig.240259]

作者简介

蔡伟南，男，硕士研究生，研究方向为三维生成。E-mail: caiweinan22@mails.ucas.ac.cn

张源奔，通信作者，男，副研究员，主要研究方向为时空信息综合处理与应用。E-mail: zhangyb@aircas.ac.cn

王宗继，男，助理研究院，研究方向为计算机视觉、计算机图形学、三维场景重建与理解。E-mail: wangzongji@aircas.ac.cn

殷煜昊，男，硕士研究生，研究方向为三维场景重建与理解。E-mail: yinyuhao25@mails.ucas.ac.cn

刘俊义，男，研究员，研究方向为信号与信息处理。E-mail: liujy004735@aircas.ac.cn